

## 영화 투자자를 위한 흥행성과 예측지표 발굴

김유진<sup>1</sup>, 권오경<sup>2</sup>

### 요약

본 연구의 목적은 영화제작 초기 단계에 크라우드 펀딩(crowd funding) 등의 방식으로 영화에 투자하는 일반 투자자의 입장에서 의미 있는 투자지표를 발굴하는 것이다. 대표적인 지도학습(supervised learning) 데이터마이닝 기법인 의사결정나무, 회귀분석, 인공신경망을 활용하여 분석한 결과 의사결정나무모형에서는 장르, 등급, 실화 및 원작의 존재 여부, 배우과위가 흥행을 결정하는 주요 요인으로 도출되었고, 회귀분석모형에서는 배우과위, 등급(12세 이상/15세 이상 관람가), 장르(SF), 실화가 존재하는 영화의 흥행성과가 높은 것으로 나타났다. 또한 영화장르와 계절(개봉시기) 간의 상호작용효과 측면에서는 봄에 개봉된 사극, 공포, 로맨스 영화와 여름에 개봉된 액션, 범죄, 사극 영화, 가을에 개봉된 액션, 범죄, 사극, 로맨스 장르 영화의 흥행성과가 좋은 것으로 나타났다. 마지막으로 인공신경망 기법에서는 12개의 입력변수를 활용하여 예측력 87.7%의 모형이 도출되었다. 본 연구의 3가지 방법론으로부터 도출된 결과를 바탕으로 투자자가 영화 산업에 투자할 경우 근거에 기반한 보다 합리적인 의사결정이 가능할 것으로 기대된다. 또한 본 연구는 투자 초기에 수집 가능한 정보만을 사용하여 영화의 흥행성과를 예측할 수 있는 지표를 발굴하였다는 측면에서 기존 연구와의 차별점을 갖는다.

주요용어 : 영화산업, 크라우드 펀딩, 의사결정나무, 다중회귀분석, 인공신경망.

### 1. 서론

2000년대 들어 한국 영화산업은 멀티플렉스 확산, 1인당 국민소득 증가, 5일 근무제 도입에 따른 여가시간 확대 등의 영향으로 크게 성장하였다. 이와 같은 한국 영화산업의 외형적인 성장과 더불어 최근 들어 국내 영화시장에서는 일반 투자자를 대상으로 한 크라우드 펀딩(crowd funding) 바람이 새롭게 불고 있다. 영화 크라우드 펀딩이란 영화사가 영화팬들에게 직접 영화를 소개하고, 일반 대중으로부터 제작비를 투자받아 팬들과 함께 영화를 만들어 가는 방식으로 2015년 7월 크라우드 펀딩법(자본시장법 개정안)이 통과된 후 2016년 1월부터 시작되었다.

이와 같이 최근 영화 크라우드 펀딩에 대한 관심이 급증한 배경에는 영화 제작자와 영화 투자자의 니즈가 부합되는 측면이 존재한다. 먼저 영화 제작사 입장에서는 크라우드 펀딩을 통해 투자자금을 모집할 수 있을 뿐만 아니라 다수의 대중 투자자가 영화 제작에 투자하고 관심을 가짐으로써 흥행을 이끄는 마케터로 활동할 것이라 기대할 수 있다. 또한 투자자 입장에서도 영화가 대중에게 친숙하고, 상대적으로 단기간 내에 수익을 실현할 수 있는 투자 대안으로 인식될 수 있다. 예를 들어 일반 대중이 창업 초기의 벤처기업 주식에 투자할 경우 기업의 기술 수준 및 향후 성장성 뿐만 아니라 비즈니스 모델의 차별성 등을 평가해야 하기 때문에 기업 외부의 이해관계자로서 기업 내부 정보의 수집 및 평가가 쉽지 않다. 반면, 영화 투자의 경우 향후 흥행 여부에 대한 판단만

<sup>1</sup>(교신저자) 07321 서울시 영등포구 의사당대로 82, 하나금융경영연구소 수석연구원.

E-mail : jazzgene@naver.com

<sup>2</sup>07328 서울시 영등포구 국제금융로6길 38, 보험연구원 연구원. E-mail : katie5022@kiri.or.kr

[접수 2017년 6월 29일; 수정 2017년 7월 27일, 2017년 8월 11일; 게재확정 2017년 8월 14일]

요구되어 상대적으로 개인 투자자 입장에서의 진입 장벽이 낮은 것으로 인식되고 있다.

한편, 영화와 같은 문화산업의 경우 수익률 변동성이 매우 커 투자자 입장에서 성과에 대한 불확실성이 큰 산업이다. 최근 10년 간 한국 영화산업의 투자수익률(ROI 기준)을 살펴보면 2008년 -43.5%에서 2014년 0.3%로 그 범위가 일반 기업의 수익성과는 비교도 되지 않는 높은 변동성을 보여왔다. 실제 영화산업에 대한 크라우드 펀딩에서도 비슷한 결과를 보였는데, 2016년 7월에 개봉한 영화 “인천상륙작전”의 경우 펀딩을 통하여 288명으로부터 5억 원의 투자를 유치하였고, 개봉 후 누적 관객수가 700만 명을 상회하면서 개인 투자자들에게 25.6%의 투자수익을 안겨 주었다. 반면 비슷한 시기에 개봉한 영화 “사냥”의 경우 289명이 투자해 목표액 3억 원을 넘기며 펀딩에 성공하였으나, 목표관객 164만 명의 절반에도 못 미치는 64만 명 정도의 관객만을 동원하면서 투자자들은 50% 가량의 투자 손실을 입었다.

실제 영화제작사가 자금을 모집함에 있어 대형투자사(기업)에게는 거액의 제작비를 지원받기 위해 제작 진행상황, 촬영본, 편집본 등을 공개하는 반면, 일반 소액 투자자에게는 줄거리 요약과 감독, 배우의 프로필 정도만 제공하는 것으로 알려져 있다. 이에 일반 소액 투자자 입장에서는 정보 비대칭이 매우 커 투자수익을 얻기가 쉽지 않다. 따라서 영화에 투자하는 개인 투자자 입장에서는 영화 제작 초기 단계에서 얻을 수 있는 한정된 영화의 정보 중 향후 흥행에 영향을 미치는 요인에 대한 관심이 클 수밖에 없다.

따라서 본 연구에서는 영화 크라우드 펀딩 투자를 고려하는 개인 투자자의 입장에서 영화 제작 단계에서 영화흥행에 가장 큰 영향을 미치는 요인이 무엇인지를 알아보고자 한다. 이에 의사결정나무, 회귀분석모형, 인공지능망 등 3가지 방법론을 활용하여 비교·분석함으로써 정보가 부족한 일반 투자자에게 유익한 투자지표를 제시하고자 한다. 본 연구의 구성은 다음과 같다. 다음 2장에서는 국내외 선행연구들에서 영화흥행과 관련한 다양한 연구에 대해 살펴본다. 3장에서는 본 논문에서 진행한 연구 방법을 소개하며, 이어지는 4장에서 분석 결과를 논의하고 마지막 5장에서 요약 및 결론을 제시한다.

## 2. 선행연구

영화산업의 수익창출 개념인 흥행에 관심을 둔 선행연구는 다양하게 진행되어 왔으며, 특히 1970년대 이후 수익에 영향을 미치는 요소들을 구조적인 차원에서 살펴보면서 그 효과성이 실증적으로 입증되었다. 먼저 Litman(1983)은 기존 연구들을 정리하여 영화산업의 흥행에 있어 중요한 요인을 크게 창조(제작)영역, 배급(유통)영역, 상영(마케팅)영역 3가지로 구분하여 제시하였다. 제작영역에 속하는 요인은 영화의 내용과 질에 관한 것으로 장르, 상영 등급, 스타파워, 제작비 등이 있으며, 배급영역의 경우 유통 배급사의 메이저 여부 및 휴일 개봉시기 등의 요인으로 구성된다. 마지막으로 상영영역은 영화의 아카데미상 수상 여부 및 평론가 및 일반인(네티즌) 평가와 관련된 것이다. Shochay(1994)는 비평가 평론, 영화 개봉관 수, 스타배우 출연 및 아카데미상 수상 여부와 더불어 경쟁의 정도가 영화흥행에 큰 영향을 끼치고 있다는 점을 확인하였다. 이후 Eliashberg, Shugan(1997), Ravid(1999) 등도 영화의 장르, 스타배우 유무, 상영등급, 제작비 규모, 배급사의 특성을 바탕으로 영화의 성공과 실패에 영향을 미치는 요인을 찾고 그 영향력을 분석한 바 있다.

한국 영화산업의 흥행성과 요인에 대한 연구는 스타파워를 분석한 Kim(1998) 이후 본격화되었으며, Yu, Kim(2002)은 한국영화의 제작과 배급단계의 흥행변수에 관해 연구하였다. Kim(2003)은 한국영화와 할리우드 영화의 흥행성과에 영향을 미친 요인을 각각 분석하여 비교하였는데 제작비, 개봉스크린 규모, 장르, 등급, 감독파워가 흥행성과에 유의미한 영향을 미쳤다는 결론을 제시하였

다. 이후 다양한 연구에서 영화 흥행요인을 찾으려는 노력이 다각도로 진행되어 왔으며, 최근 연구로 Choi et al.(2009)은 영화 흥행 결정요인을 통합적으로 분석하여 감독파워, 스타파워, 네티즌 리뷰 수가 영화 관객 수에 유의한 양의 영향을 미침을 확인하였다. 이 외에 Lee, Pyo(2009)는 한국 영화 흥행요인의 잠재적 효과에 대해 구조방정식을 이용하여 분석하였는데, 영화흥행 요소에 영화의 스타와 감독의 인지도인 영화속성 변수가 잠재변수로 영향을 주는 것으로 나타났다. 이후 Park, Song(2010)은 영화흥행에 영향을 미치는 요인으로 제작비 규모에 따라 흥행성과 요인의 영향력이 차별화되는 것을 확인하였는데, 제작비 규모가 커질수록 기존 연구에서 확인된 흥행결정 요인의 영향력이 커지는 것으로 나타났다.

기존 연구를 통해 영화 흥행성과의 요인으로 중요하게 지적되는 변수는 장르, 배급사, 등급, 배우 변수이며, 그 외 다양한 변수가 영향을 미치고 있음이 확인되었고 이들 변수의 영향력이 복잡하게 나타나고 있는 것을 알 수 있었다. 또한 영화흥행을 평가함에 있어 대부분 관객 수(box-office)를 기준으로 한국영화의 성과를 측정하였음을 알 수 있었다. 이는 영화산업의 특성상 제작사가 대부분 외부감사 대상 이하 법인으로 경영 성과가 금융감독원 전자공시시스템(DART)에 공시조차 되지 않아 연구자 입장에서 수익성과 같은 재무 정보의 접근이 용이하지 않았기 때문인 것으로 판단된다. 또한 영화 매출에 있어 작품 당 가격의 차별성이 거의 없어 관객 수 또는 매출이 흥행성과를 설명하는 대표변수로 활용됨에 있어 무리가 없어 보인다.

### 3. 연구 설계

#### 3.1. 분석대상 및 주요 연구변인

연구대상은 영화진흥위원회(이하 영진위) 통합전산망에서 데이터 수집이 가능한 2010년부터 2015년까지 한국 국적의 개봉영화 중 상업영화를 대상으로 하며, 매출액 천만 원 미만은 표본에서 제외하였다. 그 외 배우 및 감독파워, 원작 및 실화 존재 여부 등의 데이터 수집과정에서 결측치를 제외한 결과, 연구 표본은 최종 398개로 결정되었다. 한편, 영화 제작자는 투자자를 모집할 때 “영화투자제안서”와 같은 형식을 통해 제작 초기에 결정된 해당 영화의 특징에 대해 설명하고 있다. 따라서 본고에서는 제작 초기에 투자자가 획득 가능한 정보 중에서 다양한 지표를 활용하였다.

주요변수의 측정방법은 Table 1과 같다. 먼저 영화등급은 전체 관람가, 12세 관람가, 15세 관람가, 청소년 관람불가의 4등급으로 구분하였다. 장르의 경우 액션, 애니메이션, 코메디, 범죄물, 다큐멘터리, 드라마, 사극, 공포, 로맨스, SF, 기타 11가지로 분류하였다. 한편, 배우파워와 감독파워는 선행연구(Kim, 1998; Kim, 2003)의 변수 측정방법을 사용하였다. 배우파워는 대표 주인공 2인의 영화 개봉시점 전 3년 간 주연 출연작수의 합으로, 감독파워는 영화 개봉시점 전 3년 간 연출작수의 합으로 집계하였는데, 값이 클수록 인지도가 높아 흥행의 가능성이 높을 것으로 판단된다. 한편, 원작 및 실화의 존재 여부는 기존에 원작 소설, 만화 등의 시나리오나 실제 존재했던 사건을 모티브로 영화가 제작되었는지 여부를 나타내는 변수로 이 또한 개봉 전 작품에 대한 인지도와 관련 있을 것으로 보인다. 그 외 외국배우의 출연 여부는 주연배우 중 외국배우의 출연 여부를 나타내며, 영화 개봉시기 특징인 계절 및 경기는 특정 영화가 개봉된 시기의 계절과 해당 시점의 국내 소비성향을 대표하는 민간소비증가율로 측정하였다. 또한 특정 계절과 영화장르 간의 관련성이 흥행에 미치는 영향을 살펴보기 위해 개봉시점의 계절과 장르 변수간의 상호작용(interaction) 변수를 활용하였다. 마지막으로 종속변수인 영화의 흥행지표는 작품별 매출액을 정규화한 값을 사용하였다.

Table 1. Variable definition

Variable		Measurement	Source
Independent Variable	Grade(4)	Over 18, Over 15, Over 12, ALL	KOFIC
	Genre(11)	Drama, Comedy, Action, Crime, Romance, Historical, Horror, SF, Animation, Documentary, etc	KOFIC
	Actor power	Number of casts starring in the past 3 years of two leading actors	Naver
	Director power	Number of director's films in the past 3 years	Naver
	Original	Whether original scenarios such as original novels and comics exist	Naver
	Real story	Whether the actual event was present	Naver
	Sequel	Whether it is a sequel (series)	KOFIC
	3D, 4D	3D, 4D	KOFIC
	Foreign actor	Whether foreign actors appear in the main actors	KOFIC
	Release time	Season effect : Spring, Summer, Fall, Winter	KOFIC
		Economy index : Private consumption growth rate	BOK
	Season & genre	Genres released in a particular season	KOFIC
Dependent Variable	Performance	Sales	KOFIC

Note : KOFIC(Korea Film Council), BOK(The Bank of Korea)

### 3.2. 연구 설계

연구 방법론은 지도학습(supervised learning)의 대표적인 기법인 의사결정나무와 회귀분석, 인공지능 경망 분석을 활용하였다.

#### 1) 의사결정나무

의사결정나무 분석 기법이란 의사결정규칙을 나무 구조로 도표화하여 분류 및 예측을 수행하는 기법이다. 분석의 목적과 자료 구조에 따라 적절한 분리 기준(split criterion)과 정지 규칙(stopping rule)을 지정하여 종속변수와 설명변수 사이의 관계를 설명하는 규칙(rule)을 도출하는 계량적 분석 방법이다. 모형은 크게 마디(node)와 가지(branch)로 구성되어 있으며, 마디는 각 기능에 따라 구분된다. 구체적으로 마디의 시작인 뿌리마디(root node)는 나무 구조의 시작 및 분류의 대상이 되는 전체 자료를 의미한다. 반대로 끝마디(terminal node)는 나무줄기의 끝에 위치하여 모형에서 정한 정지규칙에 따라 더 이상의 분리가 일어나지 않는 마디이며, 그 외 중간마디가 존재한다. 한편 마디는 하나의 마디로부터 분리된 2개 이상의 마디인 자식마디(child node)와 자식마디의 상위 마디인 부모마디(parent node)로 구분할 수도 있다. 한편, 마디를 연결해주는 가지의 경우 뿌리마디부터 끝마디로 연결된 결과를 해석함에 있어 판단 기준을 제시한다.

의사결정나무 분석은 연속형과 범주형 변수를 모두 취급할 수 있기 때문에 범주형(이산형) 종속변수를 기준으로 마디가 분화되는 분류나무(classification tree)와 연속형 종속변수를 기준으로 마디가 분화되는 회귀나무(regression tree)로 나눌 수 있다. 분석에 활용하는 데이터의 구조 형태에 따라서 범주형 종속변수일 경우에는 목표변수의 각 범주에 속하는 빈도에 기초하여 분리기준을 정하며, 연속형 종속변수일 때에는 종속변수의 평균에 기초하여 분리한다. 이는 의사결정 알고리즘에 따라 종속변수와 설명변수 간의 상관도가 가장 높고 설명력이 가장 큰 변수를 선택하고, 선택된 변수의 범주를 조합하여 가장 높은 상관도를 보이는 조합을 찾아 모형을 분리하는 방식이다.

본 연구에서는 의사결정나무 모형을 통해 영화 창조단계에서 활용이 가능한 입력변수들과 종속변수인 영화의 흥행성과 간의 의미 있는 규칙을 찾아 영화흥행에 영향을 미치는 주요 변수들을 도

출하기 위해서 영화 매출액을 뿌리마디로 하는 의사결정나무 모형을 설계하였다.

## 2) 회귀분석

회귀분석이란 둘 또는 그 이상의 변수들 간에 존재하는 관련성을 분석하기 위하여 관측된 자료에서 이들 간의 함수적 관계식을 통계적 방법으로 추정하는 방법이다. 즉, 설명변수(independent variable)와 종속변수(dependent variable) 사이의 선형식을 구하여 설명변수의 값이 주어졌을 때 종속변수의 값을 예측하고 종속변수에 대한 설명변수의 예측력(영향력)을 분석하는 방법이다.

본 연구에서는 설명변수가 2개 이상인 다중회귀분석을 사용하는데, 영화 창조단계의 특징 변수와 영화의 흥행성과와의 관계를 알아보기 위해 영화의 등급, 장르, 배우파워(외국배우 출연 여부), 감독파워, 원작 및 실화의 존재 여부와 속편, 3D(4D) 제작 여부, 그리고 영화의 개봉시기의 특징 등을 설명변수로 하며, 영화의 흥행성과(매출액)를 종속변수로 설계하였다.

$$\begin{aligned} Sales_i = & \beta_0 + \beta_1 Grade_i + \beta_2 Genre_i + \beta_3 Actor_i + \beta_4 Director_i + \beta_5 Original_i \\ & + \beta_6 RealStory_i + \beta_7 Sequel_i + \beta_8 3D_i + \beta_9 ForeignActor_i + \beta_{10} Season_i \\ & + \beta_{11} Consumption_i + \beta_{12} Season_i * Genre_i + \varepsilon_i \end{aligned}$$

<변수 정의>

Sales = 매출액을 정규화한 값

Grade = 영화등급 더미변수(비교집단 - 전체 관람가)

Genre = 영화장르 더미변수(비교집단 - 기타)

Actor = 배우파워(최근 3년간 주연배우 출연 작품 수)

Director = 감독파워(최근 3년간 연출 작품 수)

Original = 원작 존재 여부 더미변수

Real story = 실화 존재 여부 더미변수

Sequel = 속편 여부 더미변수

3D = 3D(4D) 여부 더미변수

Foreign actor = 주연배우 중 외국배우 존재 여부 더미변수

Season = 계절 더미변수(비교집단 - 봄)

Consumption = 민간소비증가율

Season \* Genre = 4개의 계절 변수와 11개 장르의 상호작용변수

## 3) 인공신경망모형

인공신경망(ANN : artificial neural network)은 인간의 두뇌를 구성하는 신경인 뉴런과 유사한 방법으로 정보를 처리하는 구조이다. 인간의 신경망에서 뉴런들이 독립적으로 정보를 처리하되 각각 많은 연결선으로 정보를 분산하여 저장하는 것과 같이 인공신경망 모형은 정보를 병렬처리하는 동시에, 복잡한 정보들을 분산하여 처리한다. 즉, 매우 복잡한 구조를 지닌 데이터 사이의 관계나 패턴을 찾아내는 유연한 비선형 모형(flexible nonlinear model)의 하나로, 신경생리학과와의 유사성 때문에 다른 통계적 예측모형에 비해 보다 흥미롭게 여겨지고 있다. 기본적으로 모형은 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되며, 은닉층과 출력층 간의 함수관계 속에서 주요 변수에 대한 가중치가 정해진다. 이와 같은 신체 내 뉴런의 역치값처럼 학습 과정에서 원하는 값이 특정값 이상이면 시그모이드 함수(sigmoid function)에 의하여 1에 가까운 값으로, 특정

결과값 이하일 경우 0에 가까운 값으로 출력된다. 이 학습을 반복적으로 실행시키면서 인공신경망 모형의 연결강도는 최적화된다.

본 연구에서는 인공신경망 모형을 활용하여 영화의 흥행성과(매출액)를 출력변수로, 실적에 영향을 미치는 주요 요인들을 입력변수로 하는 은닉층 1개, 노드수 10개의 네트워크 구조(network architecture)를 설계하였다. 인공신경망 모형에서는 은닉층(hidden layer)을 1개로 구성하는 단층신경망(single layer neural network)이 주로 사용되는데, 은닉층 개수에 따른 모형의 성과차이가 크지 않으며, 노드(node)수는 입력변수가  $n$ 개일 때  $n/2 \sim 2n+1$ 이 최적인 것으로 알려져 있다. 입력변수의 경우 영화 제작단계에서 획득 가능한 영화의 주요 특성변수인 등급, 장르, 배우, 감독, 3D(4D) 제작 여부, 개봉시기(계절 등) 등을 활용하였으며, 출력변수는 매출액을 3분위로 구분하여 각각 상(high), 중(middle), 하(low)로 구분하여 분석하였다.

#### 4. 실증분석 결과

##### 4.1. 기술통계량 및 상관관계분석

Table 2는 주요변수의 기술통계량이다. 2010년 이후 개봉한 한국 영화의 박스오피스 매출은 최소 천만 원 대부터 최대 1,357억 원까지 매우 넓은 분포를 보였으며 평균은 111억 원으로 나타났다. 한편 매출액 중간값이 39억 원으로 평균과의 차이가 큰 것으로 나타나 상위 소수의 영화매출 규모가 매우 큰 것임을 알 수 있다. 그리고 배우과위의 평균은 7.7로 3년 평균 2인의 출연작품 수가 약 8편인 것을 알 수 있으며, 감독과위의 평균은 0.9로 3년 평균 1편에도 못 미치는 수의 작품을 연출하고 있는 것으로 나타났다. 더미변수로 처리된 원작 존재여부의 경우 평균은 0.2였고, 실화 존재여부와 속편 존재여부의 평균은 각각 0.07, 0.04로 나타났다. 분석기간 동안 민간소비증가율의 평균은 0.6%이며, 범위는 최소 -0.3%에서 최대 1.6%였다.

Table 2. Descriptive statistics

Variable	N	Mean	Median	Std.Dev.	Min	Max
Sales(₩ million)	398	11,149	3,915	18,233	10	135,748
Actor	398	7.78	8.00	4.19	0.00	24.00
Director	398	0.91	1.00	1.03	0.00	6.00
Original	398	0.23	0.00	0.42	0.00	1.00
Real_Story	398	0.07	0.00	0.26	0.00	1.00
Sequel	398	0.04	0.00	0.20	0.00	1.00
Consumption(%)	398	0.59	0.60	0.51	-0.30	1.60

Table 3은 주요변수간의 피어슨 상관관계를 나타낸 것이다. 매출액과 유의한 상관관계를 보이는 변수는 배우과위와 실화 존재여부로 이들 변수는 모두 0.1% 수준에서 유의한 양(+)의 상관관계를 보였다. 실화 존재여부 변수는 비교적 많은 변수와 유의한 상관관계를 보였는데 매출액, 배우과위, 감독과위와는 유의한 양(+)의 상관관계를 보였고, 원작 존재여부 변수와는 유의한 음(-)의 상관관계를 보였다. 감독과위 변수는 원작 존재여부, 속편 존재여부 변수와 유의한 양(+)의 상관관계를 보였다. 민간소비증가율 변수는 배우과위 변수와 유의하지만 약한 음(-)의 상관관계를 보였고, 타 변수와는 유의한 상관관계를 보이지 않았다. 한편, 상관관계(pearson correlation) 분석결과 변수 간의 다중공선성(multicollinearity)이 존재할 경우 회귀추정의 신뢰성이 저하될 수 있으므로, 본 연구에서는 분산팽창계수(VIF : variance inflation factor)를 이용하여 다중공선성을 확인하였는데, 모든 변수

에서 분산팽창계수가 10 이하로 나타나 회귀분석결과를 왜곡시킬 만큼 높지 않다고 판단하고 연구를 진행하였다.

Table 3. Correlation of factors

	Sales	Actor	Director	Original	Real_Story	Sequel	Consumption
Sales	1.000						
Actor	0.281*** (<.0001)	1.000					
Director	0.063 (0.209)	0.059 (0.241)	1.000				
Original	0.039 (0.435)	0.001 (0.987)	0.001 (0.984)	1.000			
Real_Story	0.255*** (<.0001)	0.148*** (0.003)	0.090* (0.073)	-0.152*** (0.002)	1.000		
Sequel	-0.047 (0.348)	-0.027 (0.585)	0.150*** (0.003)	0.030 (0.548)	-0.010 (0.850)	1.000	
Consumption	0.053 (0.295)	-0.094 (0.060)*	0.027 (0.598)	0.065 (0.198)	0.055 (0.277)	0.001 (0.987)	1.000

Note : Pearson correlations for the 398 samples. \*, \*\*, \*\*\* are respectively 5%, 1%, 0.1% of significant level

#### 4.2. 의사결정나무

연속형 종속변수에 해당하는 영화 흥행성적을 뿌리마디로 설정하고, 영화의 등급, 장르, 배우 및 감독파워, 원작 및 실화존재 여부, 속편 여부, 3D(4D), 외국인 주연배우 출연여부, 개봉시기의 특성(계절, 경기지표) 등과 같이 영화제작 초기에 정보획득이 가능한 영화의 특징을 설명변수로 하는 회귀나무(regression tree) 분석을 진행한 결과는 Figure 1과 같다. 분석 결과 영화의 흥행성적인 매출액을 설명하는 가장 주요한 첫 번째 분리 기준은 장르로 액션, 범죄, 사극, SF의 성과가 좋은 것으로 나타났다. 그리고 두 번째 분리 기준은 영화등급인데, 12세 이상, 15세 이상 관람가의 성과가 상대적으로 우수한 것으로 나타났으며, 그 외 하위분리 기준으로는 실화 및 원작의 존재 여부, 배우파워 등이 영화의 흥행을 결정하는 주요 요인으로 도출되었다. 의사결정나무 모형의 대표적인 장점이 다양한 설명변수 중 중요한 변수의 순위를 나타내 주는 것인데, 본 연구에서는 영화의 흥행성적에 영향을 미치는 주요변수로서 장르, 등급, 실화존재 여부, 배우파워, 원작의 존재 여부 순임을 알 수 있었다.

#### 4.3. 회귀분석

영화의 흥행성과(매출액)를 종속변수로 하는 다중회귀분석 모형의 분석 결과는 Table 4와 같다. 먼저 연구설계 과정에서 언급된 등급(4), 장르(11), 배우파워, 감독파워 등과 같은 영화의 특성과 더불어 영화의 개봉 계절(4)과 장르(11)간의 상호작용변수를 대상으로 단계적 방법(stepwise method)을 활용하여 최적의 설명변수를 추출하였다. 그 결과 전체 69개의 설명변수 중 27개의 설명변수가 선택되었다. 선택된 설명변수 27개는 Table 4에 표시되어 있으며, 그 중 19개의 변수만이 종속변수인 영화 매출에 유의한 영향을 미치는 것으로 나타났다.

먼저 의사결정나무 모형 분석 결과 가장 주요한 요인으로 꼽힌 장르의 경우 SF 장르만이 흥행과 유의미한 양(+)의 관계를 가지는 것으로 나타났다. 다음으로 배우파워와 등급(12세 이상, 15세

이상), 실화 존재 여부의 설명변수가 종속변수인 영화 매출과 유의한 양(+)의 관계를 보였다. 이를 통해 스타배우가 출연하고, 15세 및 12세 이상 관람가의 등급 속성을 가진, 실화를 바탕으로 한 영화가 흥행에 성공할 가능성이 높다고 해석할 수 있다. 즉, 이와 같은 결과는 다양한 연령층에게 공개가 가능한 동시에 적절한 수위의 영화 등급에서의 흥행성고가 높다는 것을 방증한다. 그 외 영화의 개봉시점과 관련하여 여름과 겨울의 흥행성고가 상대적으로 우수한 것으로 나타나, 영화 흥행에 있어 계절의 영향이 존재하는 것으로 해석된다. 한편, 영화장르와 계절 간의 상호효과를 살펴본 결과 봄에 개봉된 사극, 공포, 로맨스 영화와 여름에 개봉된 액션, 범죄, 사극, 가을에 개봉된 액션, 범죄, 사극, 로맨스 영화의 흥행성고가 좋은 것으로 나타났다.

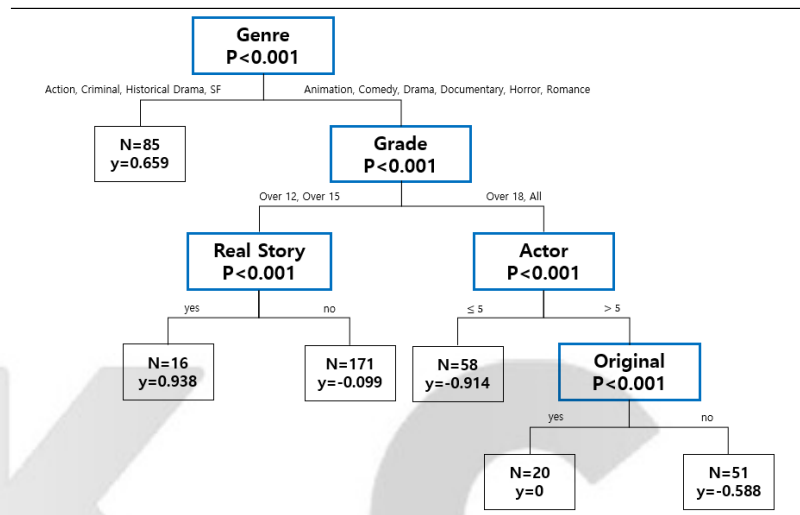


Figure 1. Result of decision tree

다만, 회귀분석 결과 기존의 흥행요소로 지적되었던 감독파워의 경우 선택적 변수선정방법(stepwise method)을 활용하여 모형에 적합한 설명변수를 추출하는 과정에서 제거되었다. 이는 기존 선행연구에서 사용한 설명변수에 속편, 실화 및 3D(4D) 여부, 계절과 장르의 상호작용변수 등이 추가되면서 흥행성고를 설명하는 최적화된 모형을 선정하는 과정에 영향을 미친 것으로 판단된다. 한편 영화매출과 감독파워의 상관관계를 살펴보면 여전히 (+)의 관계인 것을 알 수 있으며, 유의하지는 않은 것으로 나타났다.

#### 4.4. 인공신경망

본 연구에서는 영화의 다양한 속성(특징)이 성과변수인 매출(흥행) 수준에 미치는 영향을 판단하기 인공신경망 모형을 활용하였다. Figure 2는 12가지의 영화 속성이 입력층(input layer)으로 투입된 후 1개의 은닉층(hidden layer)을 거쳐 흥행성고인 출력층(output layer)으로 연결되는 과정을 보여준다. 한 처리요소의 출력이 다음 층의 어떤 처리요소로 입력되는 과정에서 입력 데이터 값들의 상대적 강도(중요성)를 나타내는 가중치가 도출되며, 가중치의 경우 임의값으로 초기화된 이후, 주어진 과거의 입력 데이터 값과 해당 출력 데이터 값들을 통해 입력 및 출력 간의 관계를 학습함으로써 적절하게 조정된다. 이와 같이 인공신경망 기법은 다층구조를 통해 입력변수와 결과변수 간의 관계를 정의하기 어려운 복잡한 비선형 데이터에 대해 높은 예측력을 가진 모형 수립이 가능하다



는 장점이 있다. 다만, 입력변수와 결과변수 간의 직접적인 상관관계를 설명하는 통계기법과 달리 인공신경망 모형은 변수 간의 인과관계를 논리적으로 설명하지 못한다는 점에서 블랙박스(black box) 모형으로 불리기도 한다.

Table 4. Results of regression analysis between factors and sales

Independent variable	Dependent variable = Sales	
	coefficient	t-value
intercept	-3.116	(-2.993)**
Genre : Drama	0.224	(1.694)
Genre : Romance	-0.416	(-1.965)
Genre : SF	1.147	(2.178)*
Grade : OVER 12	0.323	(2.291)*
Grade : OVER 15	0.464	(4.322)***
Actor	0.051	(4.216)***
Original	0.176	(1.601)
Real Story	0.780	(3.788)***
Season : Summer	2.097	(2.029)*
Season : Fall	1.883	(1.816)
Season : Winter	2.458	(2.364)*
Spring*Action	2.115	(1.903)
Spring*Animation	2.132	(1.836)
Spring*Comedy	2.104	(1.985)*
Spring*Crime	2.489	(2.205)*
Spring*Documentary	1.955	(1.603)
Spring*Drama	1.684	(1.597)
Spring*Historical	2.254	(1.990)*
Spring*Horror	2.156	(2.004)*
Spring*Romance	2.565	(2.385)*
Summer*Action	1.793	(6.493)***
Summer*Crime	0.844	(2.344)*
Summer*Historical	1.629	(4.532)***
Fall*Action	0.590	(2.221)*
Fall*Crime	1.439	(2.698)**
Fall*Historical	1.898	(3.918)***
Fall*Romance	0.605	(2.037)*
adjusted $R^2$ (%)	34.4	
F-value	(8.422)	

Note : \*, \*\*, \*\*\* are respectively 5%, 1%, 0.1% of significant level

본고에서는 학습용 표본(training data)과 검증용 표본(validation data)을 6:4로 하고, 입력노드 12개, 출력노드 1개, 은닉층 1개인 모형을 설계하였다. 학습 횟수와 은닉 노드수는 다양한 조합으로 실험한 결과 학습률이 가장 높은 학습횟수 300회, 은닉 노드수 10개가 선택되었다. 분석 결과는 Table 5와 같다. 영화흥행 수준에 대한 예측력의 정확도가 학습용 표본 97.9%, 검증용 표본 72.7%로 전체 표본의 정확도 수준은 87.7%이다. 영화의 흥행성적(매출액)인 출력변수를 3분위로 구분하여 분석한 결과를 그룹별로 살펴보면 중간그룹(middle) 대비 상위그룹(high)과 하위그룹(low)의 예측력이 더욱 높은 것으로 나타났다. 이는 표본 내 상위그룹과 중간그룹, 중간그룹과 하위그룹 간의 경계에 위치한 표본이 가진 특성에 큰 차이가 없기 때문인 것으로 판단된다. 본 결과를 통해 상위그룹과

하위그룹 간의 속성 차이가 크며, 흥행에 크게 성공할 것으로 기대되는 집단(상위)과 흥행에 실패할 것으로 예측되는 집단(하위)을 구분함에 있어 예측정확성이 높다는 것을 알 수 있다.

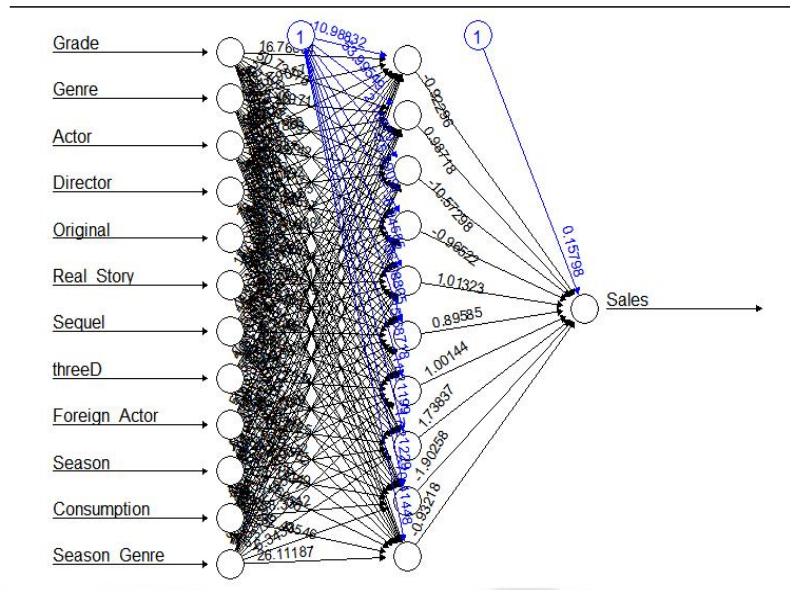


Figure 2. Results of artificial neural network

Table 5. Results of ANN

	Training set(97.9%)			Validation set(72.7%)		
	high	middle	low	high	middle	low
high	78	1	0	34	18	0
middle	1	78	3	12	37	4
low	0	0	76	4	6	46

## 5. 결론

본 연구의 목적은 영화 제작 초기 단계의 제한된 정보 하에 투자 의사결정을 하는 일반 투자자 입장에서 의미있는 투자지표를 발굴하는 것이다. 즉, 비대칭적인 정보 환경에 상황에 처한 소액 투자자가 크라우드 펀딩 등의 방법을 통해 영화산업에 투자한다고 가정할 때 “향후 흥행성파에 영향을 미치는 변수가 있을까?” 라는 질문에 답을 제시하기 위해 시작된 연구이다. 이에 대표적인 지도학습(supervised learning) 데이터마이닝 기법인 의사결정나무, 회귀분석, 인공신경망을 활용하여 분석한 결과 의사결정나무모형에서는 장르, 등급, 실화 및 원작의 존재 여부, 배우파워가 흥행을 결정하는 주요 변수로 도출되었고, 회귀분석모형에서는 배우파워 존재, 등급(12세 이상/15세 이상 관람가), 장르(SF), 실화가 존재하는 영화의 흥행성파가 높은 것으로 나타났다. 추가적으로 영화장르와 계절 간의 상호효과를 살펴본 결과 봄에 개봉된 사극, 공포, 로맨스 영화와 여름에 개봉된 액션, 범죄, 사극, 가을에 개봉된 액션, 범죄, 사극, 로맨스 영화의 흥행성파가 좋은 것으로 나타났다. 마지막으로 인공신경망 모형에서는 12개의 설명변수를 활용해 예측력 87.7%로 비교적 정확도가 높은 모형이 도출되었다.

본 연구의 결과를 바탕으로 일반 투자자가 영화에 투자하고자 할 때 다음의 3가지 방법론을 상

호 보완하여 활용할 경우 근거에 기반한 투자로 보다 합리적인 투자가 가능할 것으로 기대된다. 먼저 의사결정나무모형에서는 중요 변수의 순위를 알 수 있으며, 다음으로 다중회귀분석을 통해 주요 변수가 흥행성과에 양(+) 또는 음(-)의 영향을 미치는 정도를 상세히 살펴볼 수 있다. 마지막으로 인공지능망 모형의 경우 블랙박스 모형이라는 단점이 존재함에도 불구하고, 비교적 높은 수준의 예측력을 바탕으로 투자자 입장에서 투자대안에 대한 향후 흥행 수준을 제시한다는데 의의가 있다.

현재 한국 영화의 평균 제작기간은 약 1년 정도이며, 일반 투자자가 영화 제작 초기 단계에 투자하여 수익을 얻을 수 있는 자금 회수기간은 보통 6개월에서 1년 정도로 알려져 있다. 따라서 일반 투자자가 본 연구를 통해 발굴된 지표와 모형을 활용하여 투자 의사결정을 할 경우 투자자금 회수 시점을 기준으로 최소 6개월을 선행하는 지표를 활용할 수 있을 것이다. 더불어 본 연구는 투자 초기에 수집 가능한 정보만을 활용하여 영화의 흥행성과를 예측할 수 있다는 측면에서 기존 연구와의 차별성이 존재한다. 또한 영화 흥행성과 요인 중 기존 연구에서 고려되지 않았던 실존 존재 여부, 영화장르와 계절 간의 상호효과 변수 등을 활용하여 유의미한 결과를 도출하였다는 점에도 의의가 있다. 다만, 투자자 입장에서 흥행성과를 예측하기 위한 가장 유용한 한 가지 방법론을 추천하거나, 기존 연구에서 다루었던 기법을 융합한 새로운 모델을 제시하지 못했다는 점에서 한계점이 존재하여 향후 추가적인 연구가 필요해 보인다.

## References

- Bae, K. H. (2012). Estimation of economic value of the film industry in the national economy, *The Journal of the Korea Contents Association*, 12(9), 172-181. (in Korean).
- Back, G., Kim, K. K., Choi, S. B., Kang, C. (2015). Prediction for the films success using stylometry, *Journal of the Korean Data Analysis Society*, 17(2B), 719-728. (in Korean).
- Choi, B. H., Lee, K. J. (2013). Impact of vertical integration and revenue sharing practices on market performance in movie industry, *Journal of Korean National Economy*, 31(3), 1-27. (in Korean).
- Choi, E., Pyo, T., Park, Y., Yum, J. (2009). Integrated study in critical components that lead Korean movies to success, *Journal of the Korean Data Analysis Society*, 11(5B), 2773-2784. (in Korean).
- Chung, Y. K. (2002). *Video distribution in multi-channel era*, Korea Creative Content Agency. (in Korean).
- Eliashberg, J., Shugan, S. (1997). Film critics: Influencers or predictors?, *Journal of Marketing*, 61(2), 68-78.
- Jang, A., Park, J. Y., Park, H. W. (2015). Global communication network analysis of Busan International Film Festival (BIFF) 2014 using twitter data, *Journal of the Korean Data Analysis Society*, 17(2B), 939-949. (in Korean).
- Lee, K. J., Pyo, T. (2009). A structural equation model of success factors of Korean theatrical movies, *Journal of the Korean Data Analysis Society*, 11(6B), 3159-3169. (in Korean).
- Lim, S. J., Kim, J. S. (2011). An integrated study on determinants of performance in Korean film industry, *The Knowledge Management Society of Korea*, 12(3), 1-25. (in Korean).
- Litman, B. (1983). Predicting success of theatrical movies: An empirical study, *Journal of Popular Culture*, 16(Spring), 159-175.
- Kim, E. M. (2003). The determinants of motion picture box office performance: Evidence from movies exhibited in Korea, *Korean Journal of Journalism & Communication Studies*, 47(2), 190-220. (in Korean).
- Kim, H. J. (1998). The star power analysis of Korean movie star, *Review of Cultural Economics*, 1(1), 165-200. (in Korean).
- Korean Film Council (2016). *2014 survey on the actual condition of Korean film industry and profitability analysis of Korean movie investment*. (in Korean).
- Oh, E. H., Chon, B. S. (2007). Determinants of film critics between online users and professional reviewers, *Korean*

- Journal of Broadcasting and Telecommunication Studies*, 22(6), 267-289. (in Korean).
- Park, Y. E., Kim, S. H., Park, H. J., Lee, D. K. (2010). Exploratory study on the factors influencing the profitability of Korean movies, *Korea Business Review*, 39(2), 459-488. (in Korean).
- Park, S. H., Song, H. J. (2012). The determinants of motion picture box office performance: evidence from Korean movies released in 2011, *Journal of Social Science*, 51(1), 45-79. (in Korean).
- Ravid, A. (1999). Information, blockbusters, and stars: A study of film industry, *Journal of Business*, 72(4), 463-492.
- Sochay, S. (1994). Predicting the performance of motion pictures, *Journal of Media Economics*, 7(4), 1-20.
- Yu, S. K., Kim, S. (2002). The influence of the change of market structure in Korean film industry on the box-office record, *Journal of Media Economics and Culture*, 1(3), 58-93. (in Korean).



## Movie Performance Indicators to Predict for Investors

*YuJin Kim<sup>1</sup>, Ohkyung Kwon<sup>2</sup>*

### Abstract

The purpose of this study is to find meaningful investment indicators from the point of view of general investors who invest in movies in the early stage of making by way of crowd-funding. It is analyzed by using supervised learning techniques such as decision trees, regression, and ANN. As a result of the analysis, the decision tree model was derived as genre, grade, real story, presence of original work, and actor (star) power as the main variables that determine the box office. In the regression model, the actor power, grade (over 12 & 15 years old), genre (SF) and real story showed a significant positive value. In addition, the interaction between movie genres and seasons shows that the performances of drama, horror, romance released in the spring and action, crime, historical drama released in the summer and action, crime, historical drama, action released in the fall are good respectively. Finally, in the ANN, 12 input variables were used to derive a highly accurate model with a predictive power of 87.7%. Based on the results derived from the 3 methods of this study, it is expected that investors will make more reasonable decisions based on evidence when investing in the movie industry.

*Keywords* : movie industry, crowd funding, decision tree, multiple regression, artificial neural network.

---

<sup>1</sup>(Corresponding Author) Senior Researcher, Hana Institute of Finance, 82, Uisadang-daero, Yeongdeungpo-gu, Seoul 07321, South Korea. E-mail : jazzgene@naver.com

<sup>2</sup>Researcher, Korea Insurance Research Institute, 38, Gukjegeumyung-ro 6-gil, Yeongdeungpo-gu, Seoul 07328, South Korea. E-mail : katie5022@kiri.or.kr

[Received 29 June 2017; Revised 27 July 2017, 11 August 2017; Accepted 14 August 2017]