

# TV드라마 참여 인물의 계량 능력지표에 기반한 첫 회 시청률 상대적 우위 예측

## Predicting Relative Superiority of TV Drama First Episodes based on the Quantitative Competency Index of the Cast and Crew

주상필\*, 홍준석\*\*, 김우주\*\*\*

연세대학교 기술정책협동과정\*, 경기대학교 경영정보학과\*\*, 연세대학교 산업공학과\*\*\*

Sang Phil Ju(sangphil@yonsei.ac.kr)\*, June Seok Hong(junehong@kyonggi.ac.kr)\*\*,  
Wooju Kim(wkim@yonsei.ac.kr)\*\*\*

### 요약

TV 드라마 한 시즌 제작에 최소 수십 억 원이 투입되지만 투자 대비 효과 예측은 쉽지 않으며 참여 인력의 중요성에도 불구하고 그들에 대한 적절한 평가지표는 아직 존재하지 않는다. 그 동안 콘텐츠 평가지표로 널리 사용되어온 시청률 절대 수치는 지속 감소하고 있지만 대체할만한 지표는 아직 없는 상태다. 본 연구에서는 시청률 절대 수치가 아니라 개별 드라마 시청률 간 상대적 우위를 반응변수로 하고, 드라마 참여 인력이 과거에 획득하여 축적한 상대적 우위를 계량 능력지표화 하여 설명변수로 설계함으로써 드라마의 상대적 흥행성을 예측하는 모형을 개발하였다. 예측 모형으로는 다양한 머신러닝 알고리즘을 활용하였고 예측 성능을 높이기 위해 기존 연구에서 유용한 것으로 판명된 설명변수를 추가하여 조합하였다. 결과적으로 본 연구에서 설계한 설명변수와 기존 연구의 설명변수로부터 최적의 조합을 탐색하여 구축한 예측 모형은 84%의 높은 정분류율로 우수한 예측 성능을 보여주었다. 이렇게 본 연구에서는 TV 드라마 참여 인력 능력지표와 시청률을 활용하여 콘텐츠의 상대적 흥행성을 예측함으로써 콘텐츠 산업 전반 투자 효율화와 활성화를 촉진하려 한다.

■ 중심어 : | 드라마 | 제작진 | 출연자 | 시청률 | 예측 |

### Abstract

It is not easy to predict the return on investment in the content business, and there is no index to evaluate cast & crew. The absolute number of TV ratings is steadily declining, but there is no substitute index yet. In this study, we tried to predict the relative popularity of the drama by designing the relative superiority of the individual drama viewership as the response variable and designing the relative superiority of the drama participants as the explanatory variables. We used various machine learning algorithms and added explanatory variables that were found to be useful in previous studies. As a result, with properly combined explanatory variables, a high prediction accuracy of 84% is obtained. In this study, we intend to promote the investment efficiency of the entire contents industry by predicting the relative popularity of the contents.

■ keyword : | Drama | Crew | Cast | TV Ratings | Predicting |

## I. 서론

TV 드라마 한 시즌 제작에 수십 억 원에서 많게는 수백 억 원까지 많은 자금을 투자하지만 효과 예측이 어려워 제작 투자 의사결정이 어렵다. 광고주도 드라마 제작 과정에서 협찬 광고를 집행할 경우 드라마 흥행여부가 불투명한 상황에서 거액의 협찬 광고비를 투자하기가 쉽지 않다.

TV 시청률은 투자 의사결정에 영향을 주는 가장 중요한 요인이다. 그러나 최근에는 아래 [그림 1]처럼 전통적인 실시간 TV 이외에 스마트폰을 이용한 시청이 많아지고 있어[1] 시청률의 절대적 수치가 연도에 따라 낮아지고 있다[2]. 특히 지상파 채널의 시청률이 급격히 낮아지고 있으며 반대로 유료방송 채널과 종합편성 채널의 시청률은 점차 상승하고 있다.

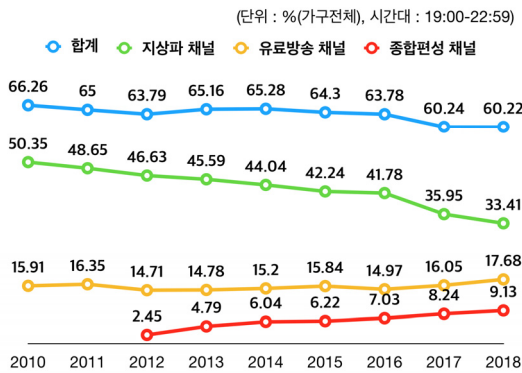


그림 1. 연간 채널별 수도권 시청률 추이[2]

그러나 유사한 기간 동안 지상파 방송사의 TV 방송 사업 매출은 큰 변화 없이 유지되고 있다[3]. [그림 2]에서 확인할 수 있는 것처럼, 2010년 약 2조 6천 억 원이었던 지상파 방송 3사 TV방송사업 합계 매출은 2017년에도 약 2조 6천 억 원으로 유사하다. 이것은 지상파 채널 TV방송사업 시장이 일종의 제로섬 게임이며 이제는 시청률의 절대 수치보다 타 방송 프로그램에 대한 상대적 우위가 시장에서의 경쟁력을 나타낼 수 있음을 암시하고 있다.

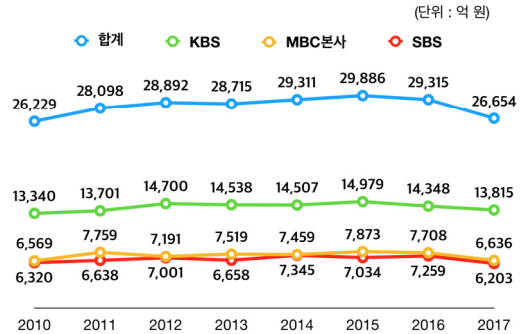


그림 2. 지상파방송사업자별 TV방송사업매출액[3]

드라마의 흥행에 영향을 주는 요인으로 가장 많이 제시되는 것이 작가, 연출자, 주요배우 등 참여인력의 능력이다. 그러나 아쉽게도 현재까지 참여인력에 대한 객관적이고 합리적인 능력 측정 지표가 거의 존재하지 않는다.

본 연구에서는 방송 드라마 참여 인력의 객관적 능력 지표를 설계하여 동시간대 방영 예정 드라마 간의 상대적 우위를 예측함으로써 드라마 제작 투자 의사 결정이나 광고주의 광고 집행 의사 결정에 도움을 주고자 한다.

## II. 선행 연구 분석

### 1. TV 시청률 예측 관련 연구

최현종(2017)은 데이터마이닝을 통해 지상파 드라마 첫 회 시청률을 예측하는 모델을 제시하였다[4]. 여기서는 드라마 첫 회 시청률에 영향을 주는 요인으로 선행 드라마의 후반 25% 시청률, 방송사, 방영시간, 방영요일과 같은 구조적인 요인과 더불어 여자배우의 과거 시청률 실적, 연출자의 과거 수상 실적과 같은 인적요인과 원작 존재 여부, 총 회차 수와 같은 것을 주요 요인으로 도출하였다. 참여 인물의 능력을 계량화 하려는 시도는 좋았으나 여자배우, 연출자와 같이 일부에 그친 것이 아쉬웠다.

배진아(2005)는 채널, 스타출연 여부, 드라마 유형, 외주제작 여부 등 드라마의 속성과 더불어 시청자 구성의 동질성, 반복 시청률 등 수용자의 속성이 드라마의

시청률에 어떤 영향을 미치는지 조사하였다[5]. 분석 결과 외주제작 여부를 제외한 나머지 요인들은 시청률에 유의미한 영향이 있는 것으로 나타났다. 그러나 출연하는 스타의 과거 성적을 계량화하지 못한 채 출연 여부만 요인으로 포함시킨 것이 아쉬웠다.

이원재(2012)는 제작사 내부에서 이루어지는 활동들이 드라마 시청률에 미치는 영향을 분석하였다[6]. 여기서 도출한 요인들은 편성요일, 시간대, 타사프로그램 경쟁력, 프로그램 길이, 시청인구 비율 등 편성관련 요인과 프로그램 회당 제작비, 스튜디오, 중계차, 편집실 등 리소스 사용 시간, 예산 집행률 등 제작비 및 리소스 관련된 요인으로 분류하였다. 그 결과로 편성과 경쟁요인 등은 시청률에 높은 영향을 미치는 것으로 분석되었으나 제작비 및 리소스 관련된 요인은 상대적으로 낮은 영향을 미치는 것으로 분석되었다. 하지만 여기서는 인적자원을 요인 후보에서 완전히 배제한 것이 좀 아쉬웠다.

강수연(2015)은 드라마 시청률에 영향을 미치는 설명변수로 작가, 연출자, 주연배우 등 인적자원뿐 아니라 프로모션, 편성시간, 방송사, 방영연도, 경쟁작, 이전작, 원작유무, 경제지표, 날씨까지 다양한 변수를 고려하였다[7]. 또한 주연배우의 경우 수상경력, 출연횟수 등을 이용해 최대한 계량화하고 프로모션을 계량화하기 위해 검색 트렌드 수치를 활용하는 등 각 변수를 상당히 계량화하는데 성공하였다. 결과도 초반 시청률, 평균 시청률 등으로 나누어 분석하였는데, 초반 시청률에 영향을 미치는 변수는 방송사, 이전작품 평균 시청률 등이었다. 그리고 평균 시청률에 영향을 미치는 변수는 초반 시청률과 포털 검색량 등이었다. 그러나 이 연구의 경우도 시청률이라는 지표가 가지는 한계에 대해 대안을 제시하지 못했다. 방송 소비 행태의 변화로 TV시청률로 측정할 수 없는 스마트폰, IPTV 등 다양한 단말을 이용한 소비가 많아지고 있으나 시청률 지표로만 분석을 한정하여 현실을 제대로 반영하지 못한 것이 아쉬웠다.

S. D. Hunter III(2016)는 시청률에 영향을 주는 독립변수로 작품의 각색여부, 작가의 과거성적을 이용하여 분석하였다[8]. 각색여부는 작품이 오리지널이면 0, 원작이 있고 각색한 것이면 1로 코딩하고, 작가의 과거

성적은 과거에 성공한 작품의 유무에 따라 0, 1, 2로 코딩하여 적용하였다. 결과는 오리지널인 경우가 시청률에 더 긍정적인 영향을 미치고 있으며, 작가의 과거 성적도 시청률에 긍정적인 영향을 미치고 있음을 발견할 수 있었다. 그러나 이 연구에서도 작가의 과거 성적을 단순히 성공한 작품의 유무에 따라 단순화 시킨 것이 아쉬움으로 남는다.

Danahera(2012)는 채널 선택에 영향을 주는 요인에 관한 연구를 통해 연간 트렌드, 하루 중 시간대, 계절, 요일이 TV프로그램 자체보다 채널 선택에 더 중요한 요인으로 작용한다는 사실을 밝혀냈다[9]. 이 연구는 편성요인이 TV 시청률 예측에 미치는 영향을 분석해냈다는 사실에 의의가 있다.

## 2. 스포츠 및 콘텐츠 결과 예측 관련 연구

오윤학(2014)은 데이터마이닝을 활용하여 프로야구 승패예측을 하는 모델을 제시하였다[10]. 2013년도 국내 프로야구 팀과 선수들의 누적 데이터를 통해 다음 경기의 승패를 예측하는 모델이었다. 승패에 영향을 미치는 주요 요인으로는 팀 평균연봉, 상대승률, 상대팀에 대한 평균자책점, 타자의 장타율, 선발투수의 승률, 이닝당 볼넷을 도출해냈다. 랜덤포레스트, 인공신경망 모형, 의사결정나무, 로지스틱 회귀분석 등 7가지 분석 기법을 적용했는데 결과적으로 랜덤포레스트를 사용하였을 때 오분류율이 가장 낮은 것으로 나타났다. 영화 <머니볼>에서처럼 타자나 투수의 능력치를 기반으로 경기 승패 예측을 시도한 흥미로운 연구였는데, 본 연구에서는 이 방식을 TV 드라마 흥행 예측에 적용하였다.

R. Sharda(2006)는 영화 박스오피스를 예측하기 위해 인공신경망 분석 방법을 이용하였다[11]. 입력 데이터로 미국영화협회의 등급기준, 경쟁정도, 스타등급, 장르, 특수효과정도, 시리즈여부, 개봉관수를 이용했고, 예측 데이터로 박스오피스 수입을 9개로 범주화하여 적용하였다. 결과로 약 75%의 적중률을 보여서 69.6%의 로지스틱 회귀분석이나 71.07%의 회귀트리 분석 방식보다 높은 성능을 보여주었다.

S. D. Hunter III(2016)는 특정 영화 작가의 직전 영화에서의 성공이 현재 영화의 성공에 매우 큰 긍정적인 영향을 미치는 것을 발견하였다[12]. 특히 특정 영화 작

가가 참여한 직전영화의 개봉 첫 주 박스오피스 성적이, 현재 영화의 개봉 첫 주 박스오피스 성적에 큰 영향을 미치는 것으로 분석되었다.

김중훈(2015)은 야구 경기 데이터를 이용하여 경기의 승패를 예측하는 분석을 딥러닝(Deep Learning) 방식을 이용하여 수행하였다[13]. 입력 데이터로는 임의의 A, B 두 팀에 대해 과거 33년간의 개인과 팀에 대한 평균자책점, 타율, 피안타수, 볼넷, 투구이닝, 삼진, 탈삼진, 승률, 홈경기여부 데이터를 이용하였다. 또한 최근 트렌드를 반영하기 위해 최근 3경기의 데이터를 추가로 입력하여 총 33개의 입력데이터를 이용하였다. 예측 데이터로는 A팀 승, B팀 승, 무승부 이렇게 3개의 카테고리로 구분하여 결과를 도출하였다. 모델로는 뉴런의 개수가 각각 8개, 9개인 2개의 히든레이어로 구성된 인공신경망을 이용하였다. 결과적으로 실제 경기 승률과 예측된 경기 승률과의 차이가 3.39%에 불과할 정도로 높은 적중률을 보였다.

J. Eliashberg(2014)는 영화 박스오피스 수입을 대본과 제작 예산만 가지고 예측하는 연구를 수행했다[14]. 작품 집필 분야에 관련한 지식과 자연어 처리 기술을 이용하여 대본에서 장르, 내용, 시맨틱 정보, bag of words 기법을 이용한 분류 등 특징을 추출하였다. 이 데이터를 kernel-based 방식의 입력 값으로 사용하여 박스오피스 수입의 성능을 측정하였다. 결과적으로 이 연구에서의 방법론이 타 방법에 비해 29%나 에러를 줄이는 높은 정확도로 박스오피스 수준을 예측하였다.

R. A. Nelson(2012)은 영화배우와 감독의 스타 파워와 영화 박스오피스 수입 간의 상관관계에 관한 연구를 수행하였다[15]. 영화 정보 사이트인 IMDB가 제공하는 Star meter 랭킹 정보를 스타의 등급으로 이용해서 9개국의 박스오피스 수입과의 관계를 분석했다. 결과적으로 특정 영화의 일반 배우 한명을 톱스타 한명으로 교체할 경우 평균 1,600만 달러가 넘는 수익이 증가하며, 세 명의 일반 배우를 세 명의 톱스타로 교체할 경우 평균 6,400만 달러가 넘는 수익의 증가를 가져온다는 사실을 밝혔다. 이 연구는 스타의 등급을 웹사이트에서 제공하는 랭킹을 이용하여 계량화 했다는 점에서는 흥미로우나 랭킹 알고리즘이 제대로 알려져 있지 않아 완

전한 설명이 어렵다는 점이 아쉬웠다.

S. Basuroy(2003)는 Variety지가 제공하는 영화 정보, 영화 비평 정보를 이용하여 영화 비평 정보가 박스오피스 수입에 미치는 영향을 연구하였다[16]. 결과적으로 긍정적, 부정적 비평 모두 개봉 후 8주간 박스오피스 수입에 영향을 미치는 것으로 나타났다. 그러나 부정적인 비평은 기간이 경과함에 따라 영향력이 감소하는 것을 발견했다. 이 연구는 박스오피스 수입 예측에 있어서 영화 비평 정보의 영향력을 밝혀낸 것에 의의가 있다.

D. Delen(2005)은 웹 기반의 영화 의사결정 시스템을 연구하였다[17]. 여기서는 경쟁정도, 장르, 슈퍼스타 여부, 특수효과, 개봉시기, 개봉관 수의 요인을 이용하여 영화를 'flop'부터 'blockbuster'까지 9개의 등급으로 예측했다. 이 연구는 기존 연구에서 도출된 다양한 요인들을 이용하고 로지스틱 회귀분석과 같은 예측 모델로 웹기반의 편리한 의사결정을 할 수 있는 시도를 보여주었다는데 의의가 있다.

### 3. 기존 연구와의 차별성

위에서 살펴본 것 같이 TV 시청률을 예측한 기존 연구들은 대부분 TV 시청률의 절대 수치를 예측하려고 하였다. 그러나 최근 스마트폰의 대중화와 지상파 이외 채널의 경쟁력 상승으로 TV 시청률의 절대 수치가 가지는 의미는 줄어들었고 오히려 드라마 간의 상대적 순위 여부가 중요해지고 있다. 또한 기존 연구에서도 제작진과 출연진이 드라마 흥행에 미치는 영향이 크다는 것은 대부분 도출해 냈으나 계량적, 체계적으로 그 영향을 지표화 한 경우는 거의 없었다.

본 연구는 프로야구 승패 예측에 사용된 방식[10]을 활용하여 제작진과 출연진에 대한 계량지표를 설계하고 그 지표를 기반으로 드라마간의 상대적 우위를 예측한다는 점에서 기존 연구와 차별성을 가진다.

## III. 예측 모형의 설계

### 1. 예측 목표

특정 시간대에 여러 개의 TV 드라마가 동시에 방영

되고 있을 경우, 2개 드라마씩 짝을 지어 시청률을 비교할 수 있다. 이때, 특정 2개 드라마 중 시청률이 높은 드라마가 '상대적 우위'에 있다고 말할 수 있다. 본 연구에서는 KBS, MBC, SBS 평일 프라임타임 밤 10시대에 동시 방영되는 미니시리즈 드라마에 대해 이러한 '상대적 우위'를 예측하려고 한다. 특히 기존 연구[4]에서 예측력이 높은 것으로 나타난 각 드라마 첫 회 시청률의 '상대적 우위'를 예측하려고 한다.

본 연구의 분석 프로세스와 예측 모형을 간략히 도식화하면 아래와 같다. 다음 절부터 각 프로세스 별 상세 내용을 설명한다.

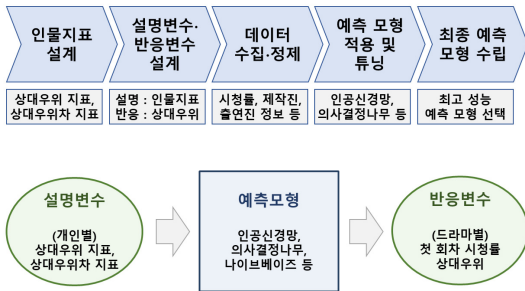


그림 3. 분석 프로세스와 예측 모형

## 2. 분석 자료

### 2.1 변수 요약

본 연구에 사용한 반응변수와 설명변수를 표로 나타내면 다음과 같다. 특히 설명변수는 본 연구를 위해 특별히 설계한 능력지표인 개인별 '상대우위 지표 (Relative Superiority Index, RSI)'와 개인별 '상대우위차 지표 (Relative Superiority Gap Index, RSGI)'를 사용하였다.

표 1. 반응변수

변수	값 및 설명
상대적우위	동시 방영되는 두 드라마 시청률을 비교하여 시청률 높은 쪽에 '1'을 부여하고, 낮거나 같은 쪽에 '0'을 부여

표 2. 설명변수

순번	변수		설명
1	A 드라마 제작진 출연진	상대 우위 (RSI)	작가
2			작가
3			출연자1
4			출연자2
5		출연자3	
6		출연자4	
7		상대 우위차 (RSGI)	작가
8			작가
9			출연자1
10			출연자2
11		출연자3	
12		출연자4	
13	B 드라마 제작진 출연진	상대 우위 (RSI)	작가
14			작가
15			출연자1
16			출연자2
17		출연자3	
18		출연자4	
19		상대 우위차 (RSGI)	작가
20			작가
21			출연자1
22			출연자2
23		출연자3	
24		출연자4	

반응변수와 설명변수에 대한 상세 사항은 다음 항에서 설명한다.

### 2.2 반응변수 설계

반응변수는 특정 드라마의 첫 회 시청률을 동시간대 방영중인 타 채널 드라마의 시청률과 비교한 결과로 정의하였다. 이때 해당 드라마의 첫 회 시청률이 타 채널 드라마의 시청률 보다 높으면 '상대적 우위'가 있다고 판단하여 '1'을 부여하고, 타 채널 드라마의 시청률보다 낮거나 같으면 '상대적 우위'가 없다고 판단하여 '0'을 부여하였다. KBS, MBC, SBS 3개 방송사가 동시간대 모두 드라마를 방영한다고 가정하면, 하루에 총 3개의 대결이 동시에 펼쳐지는 것으로 볼 수 있다. (KBS vs. MBC, MBC vs. SBS, SBS vs. KBS)

표 3. 첫 회 시청률 비교 예시

방영일	첫 회 드라마 (시청률)	기존 방영 드라마 (시청률)	
		타 채널1	타 채널2
2016.7.6.	KBS <함부로 애틀하게> (12.5%)	MBC <운빨 로맨스> (6.6%)	SBS <원티드> (7.0%)
2016.7.20.	MBC <W(더블유)> (8.6%)	KBS <함부로 애틀하게> (12.9%)	SBS <원티드> (5.4%)

[표 3]에서 보면 MBC <운빨 로맨스>와 SBS <원티드>가 방영중인 2016년 7월 6일에 KBS <함부로 애틀하게>가 12.5%의 시청률로 첫 회를 시작함으로써 기존 방영 2개 드라마 시청률 보다 상대적 우위에 있었다. 또 2016년 7월 20일에는 MBC <W(더블유)>가 시청률 8.6%로 첫 회를 시작하면서 KBS <함부로 애틀하게>보다는 낮았지만, SBS <원티드>보다는 상대적 우위에 있었던 것을 알 수 있다. 이것을 정리하면 다음과 같다.

표 4. 첫 회 시청률 반응변수 예시

방영일	첫 회 드라마	기존 방영 드라마	반응변수 (첫 회 드라마 상대적 우위)
2016.7.6.	KBS <함부로 애틀하게>	MBC <운빨 로맨스>	1
2016.7.6.	KBS <함부로 애틀하게>	SBS <원티드>	1
2016.7.20.	MBC <W(더블유)>	KBS <함부로 애틀하게>	0
2016.7.20.	MBC <W(더블유)>	SBS <원티드>	1

### 2.3 설명변수 설계

설명변수 중 우선 개인별 '상대우위 지표(RSI)' 계산 방법은 다음과 같다. 특정 드라마의 특정 회 차 시청률을, 동시간대 방영한 타 채널 드라마의 해당 회 차 시청률과 비교하여 상대적 우위 여부를 결정한다. 상대적 우위인 드라마에 참여한 연출자, 작가, 주요배우 4인은 모두 승점 1을 얻는다. 상대적 열위인 드라마에 참여한 연출자, 작가, 주요배우 4인은 승점을 얻지 못한다. 만약 시청률이 같을 경우 양 드라마에 참여한 모든 연출자, 작가, 주요배우는 승점 0.5를 얻는다. 이렇게 계산된 개인별 승점을 누적 합산한 후 개인별 총 참여 회차 수로 나누어 개인별 '상대우위 지표'를 계산한다. 양 드라마간의 비교이므로 양 드라마에 참여하는 총 12명의 '상대우위 지표'를 설명변수에 적용한다. 단, 예측하고

자하는 특정 방영일 전까지의 승점과 참여 회차 수만 누적 계산하여 적용한다.

$$\text{상대우위 (RSI)} = \frac{\text{승점 누적합계}}{\text{총 참여 드라마 회차수}} \quad (1)$$

$$(0 \leq \text{상대우위 (RSI)} \leq 1)$$

표 5. 동시간대 특정 회차 시청률 비교

방영일	KBS 드라마 시청률	MBC 드라마 시청률	SBS 드라마 시청률
2016.7.21.	11.1%	9.5%	6.5%
2016.7.27.	8.1%	12.9%	6.0%

표 6. 각 채널 드라마 정보

채널	제목	연출자	작가	주요배우 4인
KBS	함부로 애틀하게	박현석	이경희	김우빈, 배수지, 임주환, 임주은
MBC	W (더블유)	정대윤	송재정	이종석, 한효주, 김의성, 이태환
SBS	원티드	박용순	한지완	김아중, 지현우, 엄태웅, 박해준

위 [표 5]와 [표 6]에서 예시로 든 개인별 '상대우위 지표'를 계산해본다. 2016년 7월 21일에 KBS에서는 <함부로 애틀하게>가 방영되고 있었고 동시간대 MBC에서는 <W(더블유)>가, 동시간대 SBS에서는 <원티드>가 방영되고 있었다. KBS vs. MBC비교에서는 KBS <함부로 애틀하게>가 11.1%로 MBC <W(더블유)> 9.5%에 상대적 우위가 있으므로 KBS <함부로 애틀하게>에 참여한 연출자, 작가, 주요배우 4인 모두 1점을 승점으로 획득한다. 비슷한 방법으로 MBC vs. SBS 비교에서는 MBC <W(더블유)> 참여자들이 모두 승점 1점을 획득하게 되고, SBS vs. KBS 비교에서는 KBS <함부로 애틀하게> 참여자들이 모두 승점 1점을 획득하게 된다. 2016년 7월 21일 하루에 대해서만 '상대우위 지표'를 계산해보면 KBS <함부로 애틀하게>에 참여한 연출자, 작가, 주요배우 4인은 모두 각각 2점을 획득했고, MBC <W(더블유)>에 참여한 연출자, 작가, 주요배우 4인은 모두 각각 1점을 획득했고, SBS <원티드>에 참여한 연출자, 작가, 주요배우 4인은 승점을 획득하지 못했다. 각 드라마에 참여한 개인별 '상대우위 지표'는 다음과 같다.

표 7. 각 참여드라마 개인별 '상대우위 지표(RSI)' (2016.7.21. 하루)

채널	참여드라마	승점	회차수	RSI
KBS	함부로 애뜻하게	2	2	1.0
MBC	W(더블유)	1	2	0.5
SBS	원티드	0	2	0.0

개인별 '상대우위차 지표(RSGI)' 계산은 다음과 같다. 위에서 정의한 개인별 '상대우위 지표'를 계산하는 과정에서 상대적 순위일 경우 시청률 차이를 계산한다. 시청률 차이는 절대수치를 사용하지 않고 [표 8]과 같이 실제 시청률 대비 비율을 5단계로 범주화하여 사용한다. 이렇게 범주화한 시청률 차이 수치를 합산한 후 개인별 총 참여 회차 수로 나누어 개인별 '상대우위차 지표'를 계산한다.

$$\text{상대우위차}(RSGI) = \frac{\text{시청률차이 누적합계}}{\text{총 참여 드라마 회차수}} \quad (2)$$

$$(0 \leq \text{상대우위차}(RSGI))$$

예를 들어 [표 5][표 6]의 2016년 7월 21일에는 KBS <함부로 애뜻하게> 시청률이 11.1% 였고, 동시간대 대결한 SBS <원티드> 회차 시청률이 6.5%였다. <함부로 애뜻하게>에 상대적 우위가 있으므로 <함부로 애뜻하게>에 참여한 인물들만 '상대우위차 지표' 점수를 얻게 된다. 두 드라마 간 시청률 차이는 4.6%이고 <함부로 애뜻하게>의 시청률은 11.1%이므로 시청률 차이 비율은 41.4%(=4.6/11.1)가 된다. 이것을 [표 8]의 범주에서 찾아보면 2점 이므로 KBS <함부로 애뜻하게>에 참여한 연출자, 작가, 주요배우 4인은 모두 각각 2점을 획득하게 된다. 이 개인별 시청률차이 점수를 누적 합산하고 총 회차 수로 나누어 '상대우위차 지표'를 계산한다. 아래 [표 9]에서 2016년 7월 21일 하루에 대한 '상대우위차 지표'를 계산해 놓았다. '상대우위 지표'와 마찬가지로 양 드라마간의 비교이므로 양 드라마에 참여하는 총 12명에 대한 '상대우위차 지표'를 모두 설명변수에 적용한다. 또한 예측을 하려하는 특정 방영일 전까지만 누적 계산하여 적용한다.

표 8. 시청률 차이 범주화 방법

시청률 대비 시청률차 비율	시청률 차이 범주
75% 이상	4
50%이상 - 75%미만	3
25%이상 - 50%미만	2
0%이상 - 25%미만	1
0%	0

표 9. 각 참여드라마 개인별 '상대우위차 지표(RSGI)' (2016.7.21.하루)

채널	참여드라마	시청률차	회차수	RSGI
KBS	함부로 애뜻하게	3	2	1.5
MBC	W(더블유)	2	2	1.0
SBS	원티드	0	2	0.0

### 3. 예측 분석 방법

#### 3.1 분석 도구

분류 예측에 사용하는 다양한 머신러닝 분석 방법을 사용했다. 그 중에서 본 연구에서는 인공신경망, 로지스틱회귀분석, 의사결정나무, 랜덤포레스트, 서포트벡터 머신, 나이브베이즈 모델을 사용했다. 분석에 사용한 머신러닝 라이브러리는 Python 기반의 scikit-learn을 사용했으며, 작업 도구로는 웹브라우저 기반의 jupyter notebook을 사용했다.

#### 3.2 자료 수집 및 데이터 구분

2010년 1월 4일 - 2018년 12월 27일까지 KBS, MBC, SBS 월, 화, 수, 목 밤 10시대에 방영한 드라마의 시청률과 각 드라마에 참여한 메인 연출자 1인, 메인 작가 1인, 주요배우 4인에 대한 인적 정보를 수집했다. 자료는 위키피디아와 네이버에서 수집하여 교차 검증하였다. 총 274개 드라마에 대한 5,293개 회차별 시청률, 191명의 연출자, 218명의 작가, 507명의 배우 데이터를 수집했다.

수집한 데이터를 정제하여 분석용 총 479개 샘플을 확보하였다. 그러나 2013년 이전 샘플은 설명변수가 제대로 축적되지 않아 거의 0인 경우가 대부분이어서 2013년 이후의 330개 샘플만 모델 학습과 검증에 사용하였다. 330개의 약 70%인 초반 230개 샘플을

training에 사용하였고, 이후 약 15%인 50개 샘플을 validation에 사용하였으며, 나머지 가장 최근 15%인 50개 샘플을 test에 사용하였다.

### 3.3 성능 지표

각 모델별로 예측 값과 실제 값을 비교하여 True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)를 구하고, 정분류율 (accuracy), 정확도 (precision), 재현율 (recall), F1 score를 계산하여 성능을 측정하였다. 아래 [표 10]과 수식 (3), (4), (5), (6)에 그 정의를 설명해 놓았다.

표 10. Confusion Matrix

실제값 \ 예측값	1	0
1	TP	FN
0	FP	TN

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{(Precision + Recall)} \quad (6)$$

## IV. 연구 결과

### 1. 참여인력 능력지표 설명변수 적용 분석

#### 1.1 인공신경망 모델

scikit-learn의 MLPClassifier를 사용했다. 과적합 현상을 방지하고 시스템의 복잡도를 줄이면서도 예측력을 높이기 위해 히든레이어는 2개를 적용했다. 히든레이어당 최적 노드 수를 구하기 위해 validation 세트에 대한 정분류율(accuracy)이 가장 높은 히든레이어당 노드 수를 적용하였다. validation 세트에 대한 정분류율이 동일한 경우에는 training 세트에 대한 정분류율이 더 높은 노드 수를 적용했다.

표 11. 노드 수별 비교 (상위 3개만 표시)

1 <sup>st</sup> 히든레이어 노드 수	2 <sup>nd</sup> 히든레이어 노드 수	validation 정분류율	training 정분류율
1	46	0.78	0.84
3	32	0.78	0.83
6	37	0.76	0.97

[표 11]에서 볼 수 있듯이 validation 세트 정분류율은 노드 수가 (1, 46)인 경우와 (3, 32)인 경우가 동일하지만 training 세트 정분류율에서 (1, 46)인 경우가 더 높아서 (1, 46)을 히든레이어 노드 수로 채택하였다.

이 모델을 training 세트로 학습 시키고 test 세트에 적용한 결과는 아래 [표 12]와 같다.

표 12. Confusion Matrix (인공신경망)

실제값 \ 예측값	1	0
1	16	12
0	1	21
정분류율	0.74	
정확도	0.94	
재현율	0.57	

정분류율이 0.74, 정확도는 0.94로 높은 편이지만 재현율은 0.57로 낮은 편이다.

#### 1.2 로지스틱회귀분석 모델

scikit-learn의 LogisticRegression을 사용했다. 이 모델을 training 세트로 학습시키고 test 세트에 적용한 결과는 아래와 같다.

표 13. Confusion Matrix (로지스틱 회귀분석)

실제값 \ 예측값	1	0
1	17	11
0	2	20
정분류율	0.74	
정확도	0.89	
재현율	0.61	

인공신경망 모델과 유사한 결과를 보여준다. 정분류율이 0.74로 인공신경망 모델과 같다. 인공신경망모델에 비해 정확도는 0.89로 다소 떨어지지만 재현율은 0.61로 다소 높다.

#### 1.3 의사결정나무 모델

scikit-learn의 DecisionTreeClassifier를 사용했다. 이 모델을 training 세트로 학습시키고 test 세트에



적용한 결과는 아래와 같다.

표 14. Confusion Matrix (의사결정나무 모델)

실제값 \ 예측값	1	0
1	12	16
0	2	20
정분류율	0.64	
정확도	0.86	
재현율	0.43	

정분류율이 0.64, 정확도는 0.86, 재현율은 0.43으로 인공신경망 모델이나 로지스틱회귀분석 모델보다는 성능이 떨어진다.

### 1.4 랜덤포레스트 모델

scikit-learn의 RandomForestClassifier를 사용했다. 이 모델을 training 세트로 학습시키고 test 세트에 적용한 결과는 아래와 같다.

표 15. Confusion Matrix (랜덤포레스트 모델)

실제값 \ 예측값	1	0
1	15	13
0	1	21
정분류율	0.72	
정확도	0.94	
재현율	0.54	

정분류율이 0.72로 의사결정나무 모델 보다는 높지만 인공신경망 모델이나 로지스틱회귀분석에 비해 다소 떨어진다. 정확도는 0.94로 높은 편이지만 재현율이 0.54로 낮다.

### 1.5 서포트벡터머신 모델

scikit-learn의 svm.SVC를 사용했다. 이 모델을 training 세트로 학습시키고 test 세트에 적용한 결과는 아래와 같다.

표 16. Confusion Matrix (서포트벡터머신 모델)

실제값 \ 예측값	1	0
1	17	11
0	2	20
정분류율	0.74	
정확도	0.89	
재현율	0.61	

정분류율이 0.74로, 정확도는 0.89, 재현율은 0.61으로 로지스틱회귀분석과 일치한다. 인공신경망 모델과도 정분류율은 일치하지만 정확도가 다소 낮고, 재현율은 다소 높다.

### 1.6 나이브베이즈 모델

scikit-learn의 GaussianNB를 사용했다. 이 모델을 training 세트로 학습시키고 test 세트에 적용한 결과는 아래와 같다.

표 17. Confusion Matrix (나이브베이즈 모델)

실제값 \ 예측값	1	0
1	23	5
0	4	18
정분류율	0.82	
정확도	0.85	
재현율	0.82	

정분류율이 0.82로 다른 모델에 비해 가장 높다. 정확도는 0.85로, 인공신경망이나 로지스틱회귀분석 모델에 비해 다소 낮지만 재현율이 0.82로 타 모델에 비해 상당히 높다.

### 1.7 결과 요약

6개 분석 모델의 결과를 아래와 같이 표로 요약하였다. 나이브베이즈 모델로 예측한 경우가 가장 높은 정분류율, 고르게 높은 정확도, 재현율을 보여주었다.

표 18. 분석 모델별 결과 비교

구분	정분류율	정확도	재현율	F1값
인공신경망	0.74	0.94	0.57	0.71
로지스틱회귀	0.74	0.89	0.61	0.72
의사결정나무	0.64	0.86	0.43	0.57
랜덤포레스트	0.72	0.94	0.54	0.68
서포트벡터머신	0.74	0.89	0.61	0.67
나이브베이즈	<b>0.82</b>	<b>0.85</b>	<b>0.82</b>	<b>0.84</b>

## 2. 기존 연구 설명변수 적용 분석

### 2.1 분석 방법

대부분의 기존 연구는 본 연구처럼 상대적인 시청률 우위를 예측한 것이 아니라 각 드라마의 절대적인 시청

를 수치를 예측한 것이어서 본 연구와 완전한 비교는 어렵다. 그러나 본 연구의 성능을 가늠하기 위해 가장 최근에 연구된 기존 연구 자료에서 유효한 것으로 판명된 설명변수를 수집하여 본 연구의 반응변수에 대해 예측력을 검증해보았다.

최현중 등(2017)은 드라마 첫회 시청률 예측에 유효한 설명변수로 다음과 같은 것을 제시하였다[4].

표 19. 기존 논문의 유효한 설명변수

순번	변수	설명	사용
1	방송사	KBS1/KBS2/MBC/SBS	△
2	방영요일	일일/월화/수목/주말	×
3	방영시간	08/19/20/21/22	×
4	원작유무	원작을 각색한 드라마인지 여부	○
5	총 회차수	16회차/20회차 등	○
6	이전시청률	이전드라마의 후반 25% 회차에 대한 평균 시청률	○
7	여자배우 전작시청률	주요배우 4인 중 여자배우의 전작 드라마 첫회 평균 시청률 (5년간)	○
8	연출자 수상	연출자의 지난 5년간 수상 수	○

본 연구는 KBS2, MBC, SBS에서 월화/수목 밤 10시대에 방영되는 드라마에 대해서만 분석했으므로 [표 19]에서 2, 3번 변수는 생략하고, 1번 변수에서 KBS1은 생략하여 총 6개의 설명변수를 가지고 본 연구의 반응변수를 예측해보았다.

### 2.2 분석 결과

1절 연구와 동일하게 6개의 머신러닝 모델에 대해 분석을 수행했다. 결과적으로 나이브베이지 모델의 정분류율이 아래와 같이 가장 높게 나왔다.

표 20. 분석 모델별 결과 비교

구분	정분류율	정확도	재현율	F1값
인공신경망	0.58	0.82	0.32	0.46
로지스틱회귀	0.54	0.86	0.21	0.34
의사결정나무	0.62	0.71	0.54	0.61
랜덤포레스트	0.62	0.91	0.36	0.51
서포트벡터머신	0.50	0.80	0.14	0.24
나이브베이지	<b>0.66</b>	<b>0.87</b>	<b>0.46</b>	<b>0.60</b>

하지만 나이브베이지 모델의 정분류율도 0.66으로 위 1.7항에서 분석한 참여인력 능력지표 설명변수에서 얻은 정분류율 0.82에 비하면 매우 낮았다. 정확도는 0.87로 나쁘지 않았으나 재현율이 0.46으로 매우 낮았고 결과적으로 F1 Score도 0.60으로 낮게 나왔다.

## 3. 전체 설명변수 적용 분석

### 3.1 분석 방법

본 연구에서 새로 설계한 참여인력 능력지표 설명변수와 기존 연구에서 첫 회 시청률 예측에 유효한 것으로 판명된 설명변수를 모두 더하여 반응변수를 예측해보았다. 즉, [표 2]에서 제시한 본 연구의 설명변수에 [표 19]에서 제시한 기존 연구 설명변수를 모두 더하여 설명변수에 적용하고 반응변수를 예측했다. 분석 모델로는 1절 연구와 동일하게 6개의 분석 모델을 사용하였다.

### 3.2 분석 결과

1절에서 수행한 분석과 동일하게 6개의 분석 모델로 예측을 해본 결과 아래와 같이 나이브베이지 모델의 경우가 0.78의 정분류율을 보여 가장 좋은 예측력을 보여주었다. 그러나 1.7항 참여인력 능력지표 설명변수만 적용한 경우의 정분류율인 0.82에 비해서는 오히려 성능이 하락한 것을 알 수 있다.

표 21. 분석 모델별 결과 비교

구분	정분류율	정확도	재현율	F1값
인공신경망	0.67	0.68	0.55	0.61
로지스틱회귀	<b>0.74</b>	<b>0.89</b>	<b>0.61</b>	<b>0.72</b>
의사결정나무	<b>0.74</b>	<b>0.89</b>	<b>0.61</b>	<b>0.72</b>
랜덤포레스트	0.64	0.78	0.50	0.61
서포트벡터머신	0.52	0.83	0.18	0.29
나이브베이지	<b>0.78</b>	<b>0.81</b>	<b>0.79</b>	<b>0.80</b>

이 결과는 [표 19]의 설명변수 중에서 첫 회 시청률 자체를 예측하는 기존 연구에서는 유효한 것으로 판명된 설명변수이지만, 첫 회 시청률의 상대적 우위를 예측하는 본 연구에서는 유효하지 않은 설명변수가 있을 수도 있다는 것을 암시한다고 볼 수 있다.

#### 4. 최적 설명변수 적용 분석

##### 4.1 분석 방법

이번에는 기존 연구[4]에서 사용된 설명변수 중 본 연구에서도 좋은 예측 성능을 보이는 일부 설명 변수만 선택하여 [표 2]의 참여인력 능력지표 설명변수와 합한 후 반응변수 예측 성능을 분석해 보았다. [표 21]에서 높은 예측 성능을 보여 준 나이브베이즈 모델, 로지스틱회귀분석 모델, 의사결정나무 모델에 대해서만 진행했다. 각 분석 모델에 대해 validation 세트에 대한 정분류율이 가장 높은 설명변수의 조합을 골라낸 결과는 다음 [표 22][표 23][표 24]와 같다.

표 22. 로지스틱회귀분석에서 선택된 설명변수

순번	변수	설명
1	방송사	KBS1/KBS2/MBC/SBS
2	원작유무	원작을 각색한 드라마인지 여부
3	이전시청률	이전드라마의 후반 25% 회차에 대한 평균 시청률
4	여자배우 전작시청률	주요배우 4인 중 여자배우의 전작 드라마 첫 회 평균 시청률 (5년간)
5	연출자 수상	연출자의 지난 5년간 수상 수

표 23. 의사결정나무 모델에서 선택된 설명변수

순번	변수	설명
1	방송사	KBS1/KBS2/MBC/SBS
2	원작유무	원작을 각색한 드라마인지 여부
3	연출자 수상	연출자의 지난 5년간 수상 수

표 24. 나이브베이즈 모델에서 선택된 설명변수

순번	변수	설명
1	방송사	KBS1/KBS2/MBC/SBS
2	원작유무	원작을 각색한 드라마인지 여부
3	이전시청률	이전드라마의 후반 25% 회차에 대한 평균 시청률

##### 4.2 분석 결과

각 분석 모델에 대해 위 [표 22][표 23][표 24]에서 선택된 기존 연구의 설명변수와 [표 2]의 참여인력 능력지표 설명변수를 모두 합하여 반응변수를 예측한 결과를 아래 표로 정리해 보았다.

표 25. 분석 모델별 예측 결과 비교

구분	정분류율	정확도	재현율	F1값
로지스틱회귀	0.70	0.88	0.54	0.67
의사결정나무	0.66	0.82	0.50	0.62
나이브베이즈	<b>0.84</b>	<b>0.88</b>	<b>0.82</b>	<b>0.85</b>

나이브베이즈 모델로 예측한 정분류율이 0.84로 가장 높았으며 정확도와 재현율도 로지스틱회귀분석이나 의사결정나무 모델에 비해 고르게 높았다. 이 수치는 위 1.7항 [표 18]에서 확인했던 참여인력 능력지표 설명변수로만 예측한 정분류율 0.82보다도 높아 전체 분석 모델을 통틀어 정분류율이 가장 높았다. 또한 정확도와 재현율도 [표 18]의 기존 최고값보다 높거나 같은 0.88, 0.82를 보여 가장 좋은 성능을 보여주었다.

#### 5. 최종 결과 비교

위 1절에서 4절까지 분석 결과를 아래 표에서 최종적으로 비교해 보았다.

표 26. 분석 결과 최종 비교

구분	최고 성능 분석 모델	정분류율	정확도	재현율	F1값
참여인력 설명변수	나이브 베이즈	0.82	0.85	0.82	0.84
기존 연구 설명변수	나이브 베이즈	0.66	0.87	0.46	0.60
전체 설명변수	나이브 베이즈	0.78	0.81	0.79	0.80
최적 설명변수	<b>나이브 베이즈</b>	<b>0.84</b>	<b>0.88</b>	<b>0.82</b>	<b>0.85</b>

"참여인력 설명변수" 행은 위 1.7항 [표 18]에서 제시한 분석결과이고, "기존 연구 설명변수" 행은 위 2.2항 [표 20]에서 제시한 분석결과이며, "전체 설명변수" 행과 "최적 설명변수" 행은 각각 위 3.2항 [표 21], 4.2항 [표 25]에서 제시한 분석결과이다.

네 경우 모두 나이브베이즈 모델이 가장 성능이 높았으며, 특히 참여인력 능력지표 설명변수에 더하여 기존 연구 중 적절한 설명변수를 선택해 합한 "최적 설명변수" 경우가 가장 높은 정분류율과 F1 score를 보여줌으로써 가장 좋은 예측 성능을 나타냈다.

## V. 결론

본 연구 결과를 통해 드라마 참여인물의 상대적 능력 지표를 설계하여 설명변수로 적용함으로써 드라마 첫 회 시청률의 상대적 우위를 효과적으로 예측할 수 있었다.

특히, 머신러닝의 여러 모델 중 나이브베이지 모델의 예측력이 뛰어났다. 요즘 각광을 받고 있는 인공지능망 모델의 성능이 높게 나오지 않은 이유는 아마도 샘플 데이터 수가 많지 않고, 개인별 능력지표도 아직 많이 축적되지 않았기 때문으로 추정된다.

본 연구 결과는 드라마 제작자나 광고주의 투자 의사 결정에 활용 가능하다. 드라마 제작자는 방영 시기가 유사한 타 방송 드라마 작가, 연출자, 출연자 정보를 입수하여 본인이 제작하는 드라마의 첫 회 시청률이 타 드라마보다 높을지 예측할 수 있다. 예측 결과에 따라 본인 드라마의 제작진이나 출연자를 가성비 높은 인력으로 유연하게 교체 가능하므로 제작 투자 효율성을 높일 수 있다.

드라마 제작 단계에서부터 광고 투자 의사결정을 해야 하는 광고주도 본 연구결과를 활용 가능하다. 투자 대상 드라마 첫 회 시청률이 타 방송 드라마보다 높을지를 방영 전 제작 단계에서부터 예측 가능하므로 높은 광고 투자 효과를 기대할 수 있다.

이렇게 본 연구 결과를 이용해 제작자나 광고주가 캐스팅이나 광고 투자 대상 드라마 선정을 좀 더 효과적으로 할 수 있게 됨으로써 콘텐츠산업 전반적으로 투자 효율성을 향상 시킬 것으로 본다.

또한 본 연구에서는 연출자, 작가, 출연자와 같은 드라마 주요 참여인력에 대하여 계량적이고 체계적인 지표를 설계하여 제시함으로써 향후 방송 콘텐츠 분야 다양한 학문적 연구의 기초 데이터로 활용될 수 있도록 하였다.

다만 최근에는 JTBC를 비롯한 종합편성채널과 tvN을 비롯한 케이블채널 드라마도 지상파 드라마 못지않은 제작비 투자와 인기를 얻고 있으나, 그런 채널은 아직 축적된 데이터가 많지 않다. 따라서 본 연구에서는 지상파 3사 드라마에 대해서만 분석할 수밖에 없어 최근 트렌드를 완전하게 반영하지 못한 한계가 있었다. 또한 스마트폰을 이용하여 방송 프로그램을 시청하는

이용자들도 증가하고 있으나 데이터 수집에 한계가 있어 본 연구에서는 여전히 TV 시청률로만 분석 데이터를 한정해야만 한 것이 아쉬웠다.

향후 연구에서는 최근 시청률이 상승하고 있는 종합편성채널과 케이블채널까지 대상 채널을 확장하여 상대적 우위를 예측하고자 한다. 또한, TV이외에 스마트폰을 이용한 시청 데이터도 수집하여 분석함으로써 이용자의 콘텐츠 이용행태 변화를 반영하여 더 정확도 높은 연구를 진행하고자 한다.

## 참고 문헌

- [1] “2018년도 방송매체 이용행태 조사,” 방송통신위원회, pp.30-31, 2018.
- [2] <http://www.mediatoday.co.kr/?mod=news&act=articleView&idxno=144060>
- [3] “2018년도 방송시장 경쟁상황 평가,” 방송통신위원회, p.110, 2018.
- [4] 최현중, 박영선, 정수미, 김화중, “데이터 마이닝을 통한 지상파 드라마 첫 회 시청률 예측 모형 연구,” 한국정보기술학회논문지, Vol.15, No.1, pp.1-10, 2017.
- [5] 배진아, “드라마 시청률 영향 요인 분석,” 한국방송학보, Vol.19, No.2, pp.270-309, 2005.
- [6] 이원재, 이남용, 김종배, “드라마 시청률 예측모델에 대한 실증적 연구,” 한국디지털콘텐츠학회논문지, Vol.13, No.3, pp.325-334, 2012.
- [7] 강수연, 전희정, 김지혜, 송종우, “국내 드라마 시청률 예측 및 영향요인 분석,” 한국통계학회 <응용통계연구>, 제28권, 제5호, pp.933-944, 2015.
- [8] S. D. Hunter III, R. Chinta, S. Smith, A. Shamim, and A. Bawazir, “Moneyball for TV: A Model for Forecasting the Audience of New Dramatic Television Series,” *Studies in Media Communication*, Vol.4, No.2, 2016.
- [9] P. J. Danahera, T. S. Daggerb, and M. S. Smitha, “Forecasting television ratings,” *International Journal of Forecasting*, Vol.27, 2011.
- [10] 오윤학, 김한, 윤재섭, 이종석, “데이터마이닝을 활용한 한국프로야구 승패예측모형 수립에 관한 연구,” *Journal of the Korean Institute of Industrial Engineers*, Vol.40, No.1, pp.8-17, 2014.

[11] R. Sharda and D. Delen, "Predicting box-office success of motion pictures with neural networks," *Expert Systems with Applications*, Vol.30 pp.243-254, 2006.

[12] S. D. Hunter III, S. Smith, and S. Singh, "Predicting Box Office from the Screenplay: A Text Analytical Approach," *Journal of Screenwriting*, Vol.5, No.2, pp.135-154, 2016.

[13] 김종훈, 김경태, 한종기, "Deep Learning 기반 기계 학습 알고리즘을 이용한 야구 경기 Big Data 분석," *한국통신학회학술대회논문집*, pp.262-265, 2015.

[14] J. Eliashberg, S. K. Hui, and Z. J. Zhang, "Assessing Box Office Performance Using Movie Scripts: A Kernel-Based Approach," *IEEE Transactions on Knowledge and Data Engineering*, Vol.26, Issue 11, 2014.

[15] R. A. Nelson and R. Glotfelty, "Movie stars and box office revenues: An empirical analysis," *Journal of Cultural Economics*, Vol.36, No.2, pp.141-166, 2012.

[16] S. Basuroy, S. Chatterjee, and S. A. Ravid, "How critical are critical reviews? The box office effects of film critics, star power, and budgets," *Journal of Marketing*, Vol.67(October), pp.103-117, 2003.

[17] D. Delen, "Movie forecast Guru: A Web-based DSS for Hollywood managers," *Decision Support Systems*, Vol.43, Issue4, pp.1151-1170, 2005.

홍준석(June Seok Hong)

정회원



- 1997년 8월 : KAIST 테크노경영대학원(경영공학박사)
- 1999년 ~ 2002년 : 인제대학교 경영학과 조교수
- 2003년 ~ 현재 : 경기대학교 경영정보학과 교수

〈관심분야〉 : 빅데이터 분석과 딥러닝을 이용한 지능형 의사결정 시스템, 지능형 에이전트 시스템

김우주(Wooju Kim)

정회원



- 1994년 8월 : KAIST 경영학과(박사)
- 1996년 ~ 2002년 : 전북대학교 산업공학과 조교수
- 2002년 ~ 2004년 : 전북대학교 산업정보시스템 부교수
- 2004년 ~ 현재 : 연세대학교 산업

공학과 교수

〈관심분야〉 : 차세대 웹 기반 기술 및 응용, 웹 서비스 기반 기술 및 응용, 전자상거래 (EC) 및 E-Business, 경영정보 시스템 및 전문가시스템

저자소개

주상필(Sang Phil Ju)

정회원



- 2008년 2월 : KAIST 경영대학 경영정보전공 (석사)
- 2016년 3월 ~ 현재 : 연세대학교 기술정책전공 박사과정

〈관심분야〉 : 머신러닝, 콘텐츠 흥행성 예측, 투자 효과 분석